

# Tight Tradeoffs for Real-Time Approximation of Longest Palindromes in Streams

Paweł Gawrychowski<sup>1</sup>, Oleg Merkurev<sup>2</sup>, Arseny M. Shur<sup>2</sup>, and Przemysław Uznański<sup>3</sup>

<sup>1</sup>Institute of Informatics, University of Warsaw, Poland

<sup>2</sup>Institute of Mathematics and Computer Science, Ural Federal University, Ekaterinburg, Russia

<sup>3</sup>Department of Computer Science, ETH Zürich, Switzerland

October 12, 2016

## Abstract

We consider computing a longest palindrome in the streaming model, where the symbols arrive one-by-one and we do not have random access to the input. While computing the answer exactly using sublinear space is not possible in such a setting, one can still hope for a good approximation guarantee. Our contribution is twofold. First, we provide lower bounds on the space requirements for randomized approximation algorithms processing inputs of length  $n$ . We rule out Las Vegas algorithms, as they cannot achieve sublinear space complexity. For Monte Carlo algorithms, we prove a lower bound of  $\Omega(M \log \min\{|\Sigma|, M\})$  bits of memory; here  $M = n/E$  for approximating the answer with additive error  $E$ , and  $M = \frac{\log n}{\log(1+\varepsilon)}$  for approximating the answer with multiplicative error  $(1 + \varepsilon)$ . Second, we design three real-time algorithms for this problem. Our Monte Carlo approximation algorithms for both additive and multiplicative versions of the problem use  $\mathcal{O}(M)$  words of memory. Thus the obtained lower bounds are asymptotically tight up to a logarithmic factor. The third algorithm is deterministic and finds a longest palindrome exactly if it is short. This algorithm can be run in parallel with a Monte Carlo algorithm to obtain better results in practice. Overall, both the time and space complexity of finding a longest palindrome in a stream are essentially settled.

## 1 Introduction

In the streaming model of computation, a very long input arrives sequentially in small portions and cannot be stored in full due to space limitation. While well-studied in general, this is a rather recent trend in algorithms on strings. The main goals are minimizing the space complexity, i.e., avoiding storing the already seen prefix of the string explicitly, and designing real-time algorithm, i.e., processing each symbol in worst-case constant time. However, the algorithms are usually randomized and return the correct answer with high probability. The prime example of a problem on string considered in the streaming model is pattern matching, where we want to detect an occurrence of a pattern in a given text. It is somewhat surprising that one can actually solve it using polylogarithmic space in the streaming model, as proved by Porat and Porat [15]. A simpler solution was later given by Ergün et al. [6], while Breslauer and Galil designed a real-time algorithm [3]. Similar questions studied in such setting include multiple-pattern matching [4], approximate pattern matching [5], and parametrized pattern matching [10].

We consider computing a longest palindrome in the streaming model, where a palindrome is a fragment which reads the same in both directions. This is one of the basic questions concerning regularities in texts and it has been extensively studied in the classical non-streaming setting, see [1, 8, 12, 14] and the references therein. The notion of palindromes, but with a slightly different meaning, is very important in computational

biology, where one considers strings over  $\{A, T, C, G\}$  and a palindrome is a sequence equal to its reverse complement (a reverse complement reverses the sequences and interchanges  $A$  with  $T$  and  $C$  with  $G$ ); see [9] and the references therein for a discussion of their algorithmic aspects. Our results generalize to biological palindromes in a straightforward manner.

We denote by  $\text{LPS}(S)$  the problem of finding the maximum length of a palindrome in a string  $S$  (and a starting position of a palindrome of such length in  $S$ ). Solving  $\text{LPS}(S)$  in the streaming model was recently considered by Berenbrink et al. [2], who developed tradeoffs between the bound on the error and the space complexity for additive and multiplicative variants of the problem, that is, for approximating the length of the longest palindrome with either additive or multiplicative error. Their algorithms were Monte Carlo, i.e., returned the correct answer with high probability. They also proved that any Las Vegas algorithm achieving additive error  $E$  must necessarily use  $\Omega(\frac{n}{E} \log |\Sigma|)$  bits of memory, which matches the space complexity of their solution up to a logarithmic factor in the  $E \in [1, \sqrt{n}]$  range, but leaves a few questions. Firstly, does the lower bound still hold for Monte Carlo algorithms? Secondly, what is the best possible space complexity when  $E \in (\sqrt{n}, n]$  in the additive variant, and what about the multiplicative version? Finally, are there real-time algorithms achieving these optimal space bounds? We answer all these questions.

Our main goal is to settle the space complexity of  $\text{LPS}$ . We start with the lower bounds in Sect. 2. First, we show that Las Vegas algorithms cannot achieve sublinear space complexity at all. Second, we prove a lower bound of  $\Omega(M \log \min\{|\Sigma|, M\})$  bits of memory for Monte Carlo algorithms; here  $M = n/E$  for approximating the answer with additive error  $E$ , and  $M = \frac{\log n}{\log(1+\varepsilon)}$  for approximating the answer with multiplicative error  $(1+\varepsilon)$ . Then, in Sect. 3 we design real-time Monte Carlo algorithms matching these lower bounds up to a logarithmic factor. More precisely, our algorithm for  $\text{LPS}$  with additive error  $E \in [1, n]$  uses  $\mathcal{O}(n/E)$  words of space, while our algorithms for  $\text{LPS}$  with multiplicative error  $\varepsilon \in (0, 1]$  (resp.,  $\varepsilon \in (1, n]$ ) uses  $\mathcal{O}(\frac{\log(n\varepsilon)}{\varepsilon})$  (resp.,  $\mathcal{O}(\frac{\log n}{\log(1+\varepsilon)})$ ) words of space<sup>1</sup>. Finally we present, for any  $m$ , a deterministic  $\mathcal{O}(m)$ -space real-time algorithm solving  $\text{LPS}$  exactly if the answer is less than  $m$  and detecting a palindrome of length  $\geq m$  otherwise. The last result implies that if the input stream is fully random, then with high probability its longest palindrome can be found exactly by a real-time algorithm within logarithmic space.

**Notation and Definitions.** Let  $S$  denote a string of length  $n$  over an alphabet  $\Sigma = \{1, \dots, N\}$ , where  $N$  is polynomial in  $n$ . We write  $S[i]$  for the  $i$ th symbol of  $S$  and  $S[i..j]$  for its *substring* (or *factor*)  $S[i]S[i+1] \dots S[j]$ ; thus,  $S[1..n] = S$ . A *prefix* (resp. *suffix*) of  $S$  is a substring of the form  $S[1..j]$  (resp.,  $S[j..n]$ ). A string  $S$  is a *palindrome* if it equals its *reversal*  $S[n]S[n-1] \dots S[1]$ . By  $L(S)$  we denote the length of a longest palindrome which is a factor of  $S$ . The symbol  $\log$  stands for the binary logarithm.

We consider the *streaming model* of computation: the input string  $S[1..n]$  (called the *stream*) is read left to right, one symbol at a time, and cannot be stored, because the available space is sublinear in  $n$ . The space is counted as the number of  $\mathcal{O}(\log n)$ -bit machine words. An algorithm is *real-time* if the number of operations between two reads is bounded by a constant. An approximation algorithm for a maximization problem has *additive error*  $E$  (resp., *multiplicative error*  $\varepsilon$ ) if it finds a solution with the cost at least  $OPT - E$  (resp.,  $\frac{OPT}{1+\varepsilon}$ ), where  $OPT$  is the cost of optimal solution; here both  $E$  and  $\varepsilon$  can be functions of the size of the input. In the  $\text{LPS}(S)$  problem,  $OPT = L(S)$ .

A *Las Vegas algorithm* always returns a correct answer, but its working time and memory usage on the inputs of length  $n$  are random variables. A *Monte Carlo algorithm* gives a correct answer with high probability (greater than  $1 - 1/n$ ) and has deterministic working time and space.

## 2 Lower Bounds

In this section we use Yao's minimax principle [17] to prove lower bounds on the space complexity of the  $\text{LPS}$  problem in the streaming model, where the length  $n$  and the alphabet  $\Sigma$  of the input stream are specified. We denote this problem by  $\text{LPS}_\Sigma[n]$ .

<sup>1</sup>Note that (a)  $\log(1+\varepsilon)$  is equivalent to  $\varepsilon$  whenever  $\varepsilon < 1$ ; (b) the space used by the algorithms is  $\mathcal{O}(n)$  for any values of errors; (c) the multiplicative lower bound applies to  $\varepsilon > n^{-0.99}$ , and thus is not contradicting the algorithm space usage.

**Theorem 2.1** (Yao’s minimax principle for randomized algorithms). *Let  $\mathcal{X}$  be the set of inputs for a problem and  $\mathcal{A}$  be the set of all deterministic algorithms solving it. Then, for any  $x \in \mathcal{X}$  and  $A \in \mathcal{A}$ , the cost of running  $A$  on  $x$  is denoted by  $c(A, x) \geq 0$ .*

*Let  $p$  be the probability distribution over  $\mathcal{A}$ , and let  $A$  be an algorithm chosen at random according to  $p$ . Let  $q$  be the probability distribution over  $\mathcal{X}$ , and let  $X$  be an input chosen at random according to  $q$ . Then  $\max_{x \in \mathcal{X}} \mathbf{E}[c(A, x)] \geq \min_{a \in \mathcal{A}} \mathbf{E}[c(a, X)]$ .*

We use the above theorem for both Las Vegas and Monte Carlo algorithms. For Las Vegas algorithms, we consider only correct algorithms, and  $c(x, a)$  is the memory usage. For Monte Carlo algorithms, we consider all algorithms (not necessarily correct) with memory usage not exceeding a certain threshold, and  $c(x, a)$  is the correctness indicator function, i.e.,  $c(x, a) = 0$  if the algorithm is correct and  $c(x, a) = 1$  otherwise.

Our proofs will be based on appropriately chosen padding. The padding requires a constant number of fresh characters. If  $\Sigma$  is twice as large as the number of required fresh characters, we can still use half of it to construct a difficult input instance, which does not affect the asymptotics. Otherwise, we construct a difficult input instance over  $\Sigma$ , then add enough new fresh characters to facilitate the padding, and finally reduce the resulting larger alphabet to binary at the expense of increasing the size of the input by a constant factor.

**Lemma 2.2.** *For any alphabet  $\Sigma = \{1, 2, \dots, \sigma\}$  there exists a morphism  $h : \Sigma^* \rightarrow \{0, 1\}^*$  such that, for any  $c \in \Sigma$ ,  $|h(c)| = 2\sigma + 6$  and, for any string  $w$ ,  $w$  contains a palindrome of length  $\ell$  if and only if  $h(w)$  contains a palindrome of length  $(2\sigma + 6) \cdot \ell$ .*

*Proof.* We set:

$$h(c) = 11^s 01^{s-c} 10011^{s-c} 01^c 1.$$

Clearly  $|h(c)| = 2\sigma + 6$  and, because every  $h(c)$  is a palindrome, if  $w$  contains a palindrome of length  $\ell$  then  $h(w)$  contains a palindrome of length  $(2\sigma + 6) \cdot \ell$ . Now assume that  $h(w)$  contains a palindrome of length  $(2\sigma + 6) \cdot \ell$ , where  $\ell \geq 1$ . If  $\ell = 1$  then we obtain that  $w$  should contain a palindrome of length 1, which always holds. Otherwise, the palindrome contains 00 inside and we consider two cases.

1. The palindrome is centered inside 00. Then it corresponds to an odd palindrome of length  $\ell$  in  $w$ .
2. The palindrome maps some 00 to another 00. Then it corresponds to an even palindrome of length  $\ell$  in  $w$ .

In either case, the claim holds.  $\square$

For the padding we will often use an infinite string  $\nu = 0^1 1^1 0^2 1^2 0^3 1^3 \dots$ , or more precisely its prefixes of length  $d$ , denoted  $\nu(d)$ . Here 0 and 1 should be understood as two characters not belonging to the original alphabet. The longest palindrome in  $\nu(d)$  has length  $\mathcal{O}(\sqrt{d})$ .

**Theorem 2.3** (Las Vegas approximation). *Let  $\mathcal{A}$  be a Las Vegas streaming algorithms solving  $\text{LPS}_\Sigma[n]$  with additive error  $E \leq 0.99n$  or multiplicative error  $(1 + \varepsilon) \leq 100$  using  $s(n)$  bits of memory. Then  $\mathbf{E}[s(n)] = \Omega(n \log |\Sigma|)$ .*

*Proof.* By Theorem 2.1, it is enough to construct a probability distribution  $\mathcal{P}$  over  $\Sigma^n$  such that for any deterministic algorithm  $\mathcal{D}$ , its expected memory usage on a string chosen according to  $\mathcal{P}$  is  $\Omega(n \log |\Sigma|)$  in bits.

Consider solving  $\text{LPS}_\Sigma[n]$  with additive error  $E$ . We define  $\mathcal{P}$  as the uniform distribution over  $\nu(\frac{E}{2})x\$y\nu(\frac{E}{2})^R$ , where  $x, y \in \Sigma^{n'}$ ,  $n' = \frac{n}{2} - \frac{E}{2} - 1$ , and  $\$$  is a special character not in  $\Sigma$ . Let us look at the memory usage of  $\mathcal{D}$  after having read  $\nu(\frac{E}{2})x$ . We say that  $x$  is “good” when the memory usage is at most  $\frac{n'}{2} \log |\Sigma|$  and “bad” otherwise. Assume that  $\frac{1}{2}|\Sigma|^{n'}$  of all  $x$ ’s are good, then there are two strings  $x \neq x'$  such that the state of  $\mathcal{D}$  after having read both  $\nu(\frac{E}{2})x$  and  $\nu(\frac{E}{2})x'$  is exactly the same. Hence the behavior of  $\mathcal{D}$  on  $\nu(\frac{E}{2})x\$x^R\nu(\frac{E}{2})^R$  and  $\nu(\frac{E}{2})x'\$x^R\nu(\frac{E}{2})^R$  is exactly the same. The former is a palindrome of length  $n = 2n' + E + 2$ , so  $\mathcal{D}$  must answer at least  $2n' + 2$ , and consequently the latter also must contain a palindrome of length at least

$2n' + 2$ . A palindrome inside  $\nu(\frac{E}{2})x'$$$x^R\nu(\frac{E}{2})^R$  is either fully contained within  $\nu(\frac{E}{2})$ ,  $x'$ ,  $x^R$  or it is a middle palindrome. But the longest palindrome inside  $\nu(\frac{E}{2})$  is of length  $\mathcal{O}(\sqrt{E}) < 2n' + 2$  (for  $n$  large enough) and the longest palindrome inside  $x$  or  $x^R$  is of length  $n' < 2n' + 2$ , so since we have excluded other possibilities,  $\nu(\frac{E}{2})x'$$$x^R\nu(\frac{E}{2})^R$  contains a middle palindrome of length  $2n' + 2$ . This implies that  $x = x'$ , which is a contradiction. Therefore, at least  $\frac{1}{2}|\Sigma|^{n'}$  of all  $x$ 's are bad. But then the expected memory usage of  $\mathcal{D}$  is at least  $\frac{n'}{4} \log |\Sigma|$ , which for  $E \leq 0.99n$  is  $\Omega(n \log |\Sigma|)$  as claimed.

Now consider solving  $\text{LPS}_\Sigma[n]$  with multiplicative error  $(1 + \varepsilon)$ . An algorithm with multiplicative error  $(1 + \varepsilon)$  can also be considered as having additive error  $E = n \cdot \frac{\varepsilon}{1 + \varepsilon}$ , so if the expected memory usage of such an algorithm is  $o(n \log |\Sigma|)$  and  $(1 + \varepsilon) \leq 100$  then we obtain an algorithm with additive error  $E \leq 0.99n$  and expected memory usage  $o(n \log |\Sigma|)$ , which we already know to be impossible.  $\square$

Now we move to Monte Carlo algorithms. We first consider exact algorithms solving  $\text{LPS}_\Sigma[n]$ ; lower bounds on approximation algorithms will be then obtained by padding the input appropriately. We introduce an auxiliary problem  $\text{midLPS}_\Sigma[n]$ , which is to compute the length of the middle palindrome in a string of even length  $n$  over an alphabet  $\Sigma$ .

**Lemma 2.4.** *There exists a constant  $\gamma$  such that any randomized Monte Carlo streaming algorithm  $\mathcal{A}$  solving  $\text{midLPS}_\Sigma[n]$  or  $\text{LPS}_\Sigma[n]$  exactly with probability  $1 - \frac{1}{n}$  uses at least  $\gamma \cdot n \log \min\{|\Sigma|, n\}$  bits of memory.*

*Proof.* First we prove that if  $\mathcal{A}$  is a Monte Carlo streaming algorithm solving  $\text{midLPS}_\Sigma[n]$  exactly using less than  $\lfloor \frac{n}{2} \log |\Sigma| \rfloor$  bits of memory, then its error probability is at least  $\frac{1}{n|\Sigma|}$ .

By Theorem 2.1, it is enough to construct probability distribution  $\mathcal{P}$  over  $\Sigma^n$  such that for any deterministic algorithm  $\mathcal{D}$  using less than  $\lfloor \frac{n}{2} \log |\Sigma| \rfloor$  bits of memory, the expected probability of error on a string chosen according to  $\mathcal{P}$  is at least  $\frac{1}{n|\Sigma|}$ .

Let  $n' = \frac{n}{2}$ . For any  $x \in \Sigma^{n'}$ ,  $k \in \{1, 2, \dots, n'\}$  and  $c \in \Sigma$  we define

$$w(x, k, c) = x[1]x[2]x[3] \dots x[n']x[n']x[n' - 1]x[n' - 2] \dots x[k + 1]c0^{k-1}.$$

Now  $\mathcal{P}$  is the uniform distribution over all such  $w(x, k, c)$ .

Choose an arbitrary maximal matching of strings from  $\Sigma^{n'}$  into pairs  $(x, x')$  such that  $\mathcal{D}$  is in the same state after reading either  $x$  or  $x'$ . At most one string per state of  $\mathcal{D}$  is left unpaired, that is at most  $2^{\lfloor \frac{n}{2} \log |\Sigma| \rfloor - 1}$  strings in total. Since there are  $|\Sigma|^{n'} = 2^{n' \log |\Sigma|} \geq 2 \cdot 2^{\lfloor \frac{n}{2} \log |\Sigma| \rfloor - 1}$  possible strings of length  $n'$ , at least half of the strings are paired. Let  $s$  be longest common suffix of  $x$  and  $x'$ , so  $x = vcs$  and  $x' = v'c's$ , where  $c \neq c'$  are single characters. Then  $\mathcal{D}$  returns the same answer on  $w(x, n' - |s|, c)$  and  $w(x', n' - |s|, c)$ , even though the length of the middle palindrome is exactly  $2|s|$  in one of them, and at least  $2|s| + 2$  in the other one. Therefore,  $\mathcal{D}$  errs on at least one of these two inputs. Similarly, it errs on either  $w(x, n' - |s|, c')$  or  $w(x, n' - |s|, c)$ . Thus the error probability is at least  $\frac{1}{2n'|\Sigma|} = \frac{1}{n|\Sigma|}$ .

Now we can prove the lemma for  $\text{midLPS}_\Sigma[n]$  with a standard amplification trick. Say that we have a Monte Carlo streaming algorithm, which solves  $\text{midLPS}_\Sigma[n]$  exactly with error probability  $\varepsilon$  using  $s(n)$  bits of memory. Then we can run its  $k$  instances simultaneously and return the most frequently reported answer. The new algorithm needs  $\mathcal{O}(k \cdot s(n))$  bits of memory and its error probability  $\varepsilon_k$  satisfies:

$$\varepsilon_k \leq \sum_{2i < k} \binom{k}{i} (1 - \varepsilon)^i \varepsilon^{k-i} \leq 2^k \cdot \varepsilon^{k/2} = (4\varepsilon)^{k/2}.$$

Let us choose  $\kappa = \frac{1}{6} \frac{\log(4/n)}{\log(1/(n|\Sigma|))} = \frac{1}{6} \frac{1 - o(1)}{1 + \log |\Sigma| / \log n} = \Theta(\frac{\log n}{\log n + \log |\Sigma|}) = \gamma \cdot \frac{1}{\log |\Sigma|} \log \min\{|\Sigma|, n\}$ , for some constant  $\gamma$ . Now we can prove the theorem. Assume that  $\mathcal{A}$  uses less than  $\kappa \cdot n \log |\Sigma| = \gamma \cdot n \log \min\{|\Sigma|, n\}$  bits of memory. Then running  $\lfloor \frac{1}{2\kappa} \rfloor \geq \frac{3}{4} \frac{1}{2\kappa}$  (which holds since  $\kappa < \frac{1}{6}$ ) instances of  $\mathcal{A}$  in parallel requires less than  $\lfloor \frac{n}{2} \log |\Sigma| \rfloor$  bits of memory. But then the error probability of the new algorithm is bounded from above by

$$\left(\frac{4}{n}\right)^{\frac{3}{16\kappa}} = \left(\frac{1}{n|\Sigma|}\right)^{\frac{18}{16}} \leq \frac{1}{n|\Sigma|}$$

which we have already shown to be impossible.

The lower bound for  $\text{midLPS}_\Sigma[n]$  can be translated into a lower bound for solving  $\text{LPS}_\Sigma[n]$  exactly by padding the input so that the longest palindrome is centered in the middle. Let  $x = x[1]x[2]\dots x[n]$  be the input for  $\text{midLPS}_\Sigma[n]$ . We define

$$w(x) = x[1]x[2]x[3]\dots x[n/2] \underbrace{1000\dots 0}_n 1x[n/2+1]\dots x[n].$$

Now if the length of the middle palindrome in  $x$  is  $k$ , then  $w(x)$  contains a palindrome of length at least  $n+k+2$ . In the other direction, any palindrome inside  $w(x)$  of length  $\geq n$  must be centered somewhere in the middle block consisting of only zeroes and both ones are mapped to each other, so it must be the middle palindrome. Thus, the length of the longest palindrome inside  $w(x)$  is exactly  $n+k+2$ , so we have reduced solving  $\text{midLPS}_\Sigma[n]$  to solving  $\text{LPS}_\Sigma[2n+2]$ . We already know that solving  $\text{midLPS}_\Sigma[n]$  with probability  $1 - \frac{1}{n}$  requires  $\gamma \cdot n \log \min\{|\Sigma|, n\}$  bits of memory, so solving  $\text{LPS}_\Sigma[2n+2]$  with probability  $1 - \frac{1}{2n+2} \geq 1 - \frac{1}{n}$  requires  $\gamma \cdot n \log\{|\Sigma|, n\} \geq \gamma' \cdot (2n+2) \log \min\{|\Sigma|, 2n+2\}$  bits of memory. Notice that the reduction needs  $\mathcal{O}(\log n)$  additional bits of memory to count up to  $n$ , but for large  $n$  this is much smaller than the lower bound if we choose  $\gamma' < \frac{\gamma}{4}$ .  $\square$

To obtain a lower bound for Monte Carlo additive approximation, we observe that any algorithm solving  $\text{LPS}_\Sigma[n]$  with additive error  $E$  can be used to solve  $\text{LPS}_\Sigma[\frac{n-E}{E+1}]$  exactly by inserting  $\frac{E}{2}$  zeroes between every two characters, in the very beginning, and in the very end. However, this reduction requires  $\log(\frac{E}{2}) \leq \log n$  additional bits of memory for counting up to  $\frac{E}{2}$  and cannot be used when the desired lower bound on the required number of bits  $\Omega(\frac{n}{E} \log \min(|\Sigma|, \frac{n}{E}))$  is significantly smaller than  $\log n$ . Therefore, we need a separate technical lemma which implies that both additive and multiplicative approximation with error probability  $\frac{1}{n}$  require  $\Omega(\log n)$  bits of space.

**Lemma 2.5.** *Let  $\mathcal{A}$  be any randomized Monte Carlo streaming algorithm solving  $\text{LPS}_\Sigma[n]$  with additive error at most  $0.99n$  or multiplicative error at most  $n^{0.49}$  and error probability  $\frac{1}{n}$ . Then  $\mathcal{A}$  uses  $\Omega(\log n)$  bits of memory.*

*Proof.* By Theorem 2.1, it is enough to construct a probability distribution  $\mathcal{P}$  over  $\Sigma^n$ , such that for any deterministic algorithm  $\mathcal{D}$  using at most  $s(n) = o(\log n)$  bits of memory, the expected probability of error on a string chosen according to  $\mathcal{P}$  is  $\frac{1}{2^{s(n)+2}}$ .

Let  $n' = s(n) + 1$ . For any  $x, y \in \Sigma^{n'}$ , let  $w(x, y) = \nu(\frac{n}{2} - n')^R x y^R \nu(\frac{n}{2} - n')$ . Observe that if  $x = y$  then  $w(x, y)$  contains a palindrome of length  $n$ , and otherwise the longest palindrome there has length at most  $2n' + O(\sqrt{n}) = O(\sqrt{n})$ , thus any algorithm with additive error of at most  $0.99n$  or with a multiplicative error at most  $n^{0.49}$  must be able to distinguish between these two cases (for  $n$  large enough).

Let  $S \subseteq \Sigma^{n'}$  be an arbitrary family of strings of length  $n'$  such that  $|S| = 2 \cdot 2^{s(n)}$ , and let  $\mathcal{P}$  be the uniform distribution on all strings of the form  $w(x, y)$ , where  $x$  and  $y$  are chosen uniformly and independently from  $S$ . By a counting argument, we can create at least  $\frac{|S|}{4}$  pairs  $(x, x')$  of elements from  $S$  such that the state of  $\mathcal{D}$  is the same after having read  $\nu(\frac{n}{2} - n')^R x$  and  $\nu(\frac{n}{2} - n')^R x'$ . (If we create the pairs greedily, at most one such  $x$  per state of memory can be left unpaired, so at least  $|S| - 2^{s(n)} = \frac{|S|}{2}$  elements are paired.) Thus,  $\mathcal{D}$  cannot distinguish between  $w(x, x')$  and  $w(x, x)$ , and between  $w(x', x')$  and  $w(x', x)$ , so its error probability must be at least  $\frac{|S|/2}{|S|^2} = \frac{1}{4 \cdot 2^{s(n)}}$ . Thus if  $s(n) = o(\log n)$ , the error rate is at least  $\frac{1}{n}$  for  $n$  large enough, a contradiction.  $\square$

Combining the reduction with the technical lemma and taking into account that we are reducing to a problem with string length of  $\Theta(\frac{n}{E})$ , we obtain the following.

**Theorem 2.6** (Monte Carlo additive approximation). *Let  $\mathcal{A}$  be any randomized Monte Carlo streaming algorithm solving  $\text{LPS}_\Sigma[n]$  with additive error  $E$  with probability  $1 - \frac{1}{n}$ . If  $E \leq 0.99n$  then  $\mathcal{A}$  uses  $\Omega(\frac{n}{E} \log \min\{|\Sigma|, \frac{n}{E}\})$  bits of memory.*

*Proof.* Define  $\sigma = \min\{|\Sigma|, \frac{n}{E}\}$ .

Because of Lemma 2.5 it is enough to prove that  $\Omega(\frac{n}{E} \log \sigma)$  is a lower bound when

$$E \leq \frac{\gamma}{2} \cdot \frac{n}{\log n} \log \sigma. \quad (1)$$

Assume that there is a Monte Carlo streaming algorithm  $\mathcal{A}$  solving  $\text{LPS}_\Sigma[n]$  with additive error  $E$  using  $o(\frac{n}{E} \log \sigma)$  bits of memory and probability  $1 - \frac{1}{n}$ . Let  $n' = \frac{n-E/2}{E/2+1} \geq \frac{n}{E}$  (the last inequality, equivalent to  $n \geq E \cdot \frac{E}{E-2}$  holds because  $E \leq 0.99n$  and because we can assume that  $E \geq 200$ ). Given a string  $x[1]x[2] \dots x[n']$ , we can simulate running  $\mathcal{A}$  on  $0^E x[1] 0^{E/2} x[2] 0^{E/2} x[3] \dots 0^{E/2} x[n'] 0^{E/2}$  to calculate  $R$  (using  $\log(E/2) \leq \log n$  additional bits of memory), and then return  $\lfloor \frac{R}{E/2+1} \rfloor$ . We call this new Monte Carlo streaming algorithm  $\mathcal{A}'$ . Recall that  $\mathcal{A}$  reports the length of the longest palindrome with additive error  $E$ . Therefore, if the original string contains a palindrome of length  $r$ , the new string contains a palindrome of length  $\frac{E}{2} \cdot (r+1) + r$ , so  $R \geq r(E/2+1)$  and  $\mathcal{A}'$  will return at least  $r$ . In the other direction, if  $\mathcal{A}'$  returns  $r$ , then the new string contains a palindrome of length  $r(E/2+1)$ . If such palindrome is centered so that  $x[i]$  is matched with  $x[i+1]$  for some  $i$ , then it clearly corresponds to a palindrome of length  $r$  in the original string. But otherwise every  $x[i]$  within the palindrome is matched with 0, so in fact the whole palindrome corresponds to a streak of consecutive zeroes in the new string and can be extended to the left and to the right to start and end with  $0^E$ , so again it corresponds to a palindrome of length  $r$  in the original string. Therefore,  $\mathcal{A}'$  solves  $\text{LPS}_\Sigma[n']$  exactly with probability  $1 - \frac{1}{(n'(E/2+1)+E/2)} \geq 1 - \frac{1}{n'}$  and uses  $o(\frac{n'(E/2+1)+E/2}{E/2} \log \sigma) + \log n = o(n' \log \sigma) + \log n$  bits of memory. Observe that by Lemma 2.4 we get a lower bound

$$\gamma \cdot n' \log \min\{|\Sigma|, n'\} \geq \frac{\gamma}{2} \cdot n' \log \sigma + \frac{\gamma}{2} \cdot \frac{n}{E} \log \sigma \geq \frac{\gamma}{2} \cdot n' \log \sigma + \log n$$

(where the last inequality holds because of Eq.(1)). Then, for large  $n$  we obtain contradiction as follows

$$o(n' \log \sigma) + \log n < \frac{\gamma}{2} \cdot n' \log \sigma + \log n. \quad \square$$

Finally, we consider multiplicative approximation. The proof follows the same basic idea as of Theorem 2.6, however is more technically involved. The main difference is that due to uneven padding, we are reducing to  $\text{midLPS}_\Sigma[n']$  instead of  $\text{LPS}_\Sigma[n']$ .

**Theorem 2.7** (Monte Carlo multiplicative approximation). *Let  $\mathcal{A}$  be any randomized Monte Carlo streaming algorithm solving  $\text{LPS}_\Sigma[n]$  with multiplicative error  $(1+\varepsilon)$  with probability  $1 - \frac{1}{n}$ . If  $n^{-0.98} \leq \varepsilon \leq n^{0.49}$  then  $\mathcal{A}$  uses  $\Omega(\frac{\log n}{\log(1+\varepsilon)} \log \min\{|\Sigma|, \frac{\log n}{\log(1+\varepsilon)}\})$  bits of memory.*

*Proof.* For  $\varepsilon \geq n^{0.001}$  the claimed lower bound reduces to  $\Omega(1)$  bits, which obviously holds. Thus we can assume that  $\varepsilon < n^{0.001}$ . Define

$$\sigma = \min\{|\Sigma|, \frac{1}{50} \frac{\log n}{\log(1+2\varepsilon)} - 2\}.$$

First we argue that it is enough to prove that  $\mathcal{A}$  uses  $\Omega(\frac{\log n}{\log(1+\varepsilon)} \log \sigma)$  bits of memory. Since  $\log(1+2\varepsilon) \leq 0.001 \log n + o(\log n)$ , we have that:

$$\frac{1}{50} \frac{\log n}{\log(1+2\varepsilon)} - 2 \geq 18 - o(1) \quad (2)$$

and consequently:

$$\frac{1}{50} \frac{\log n}{\log(1+2\varepsilon)} - 2 = \Theta\left(\frac{\log n}{\log(1+2\varepsilon)}\right). \quad (3)$$

Finally, observe that:

$$\log(1+2\varepsilon) = \Theta(\log(1+\varepsilon)) \quad (4)$$

because  $\log 2(1+\varepsilon) = \Theta(\log(1+\varepsilon))$  for  $\varepsilon \geq 1$ , and  $\log(1+\varepsilon) = \Theta(\varepsilon)$  for  $\varepsilon < 1$ . From (3) and (4) we conclude that:

$$\log \sigma = \Theta(\log \min\{|\Sigma|, \frac{\log n}{\log(1+\varepsilon)}\}). \quad (5)$$

Because of Lemma 2.5 and equations (4) and (5), it is enough to prove that  $\Omega(\frac{\log n}{\log(1+\varepsilon)} \log \sigma)$  is a lower bound when

$$\log(1+2\varepsilon) \leq \gamma \cdot \frac{\log \sigma}{100}, \quad (6)$$

as otherwise  $\Omega(\frac{\log n}{\log(1+\varepsilon)} \log \sigma) = \Omega(\frac{\log n}{\log(1+2\varepsilon)} \log \sigma) = \Omega(\log n)$ .

Assume that there is a Monte Carlo streaming algorithm  $\mathcal{A}$  solving  $\text{LPS}_\Sigma[n]$  with multiplicative error  $(1+\varepsilon)$  with probability  $1 - \frac{1}{n}$  using  $o(\frac{\log n}{\log(1+\varepsilon)} \log \sigma)$  bits of memory. Let  $x = x[1]x[2] \dots x[n']x[n'+1] \dots x[2n']$  be an input for  $\text{midLPS}_\Sigma[2n']$ . We choose  $n'$  so that  $n = (1+2\varepsilon)^{n'+1} \cdot n^{0.99}$ . Then  $n' = \log_{(1+2\varepsilon)}(n^{0.01}) - 1 = \frac{1}{100} \frac{\log n}{\log(1+2\varepsilon)} - 1$ . We choose  $i_0, i_1, i_2, i_3, \dots, i_{n'}$  so that  $i_0 + \dots + i_d = \lceil (1+2\varepsilon)^{d+1} \cdot n^{0.99} \rceil$  for any  $0 \leq d \leq n'$ .

(Observe that for  $\varepsilon = \Omega(n^{-0.98})$  we have  $i_0 > n^{0.99}$  and  $i_1, \dots, i_d > 2n^{0.01} - 1$ .) Finally we define:

$$w(x) = \nu(i_{n'})^R x[1] \nu(i_{n'-1})^R \dots x[n'] \nu(i_0)^R \nu(i_0) x[n'+1] \nu(i_1) \dots \nu(i_{n'-1}) x[2n'] \nu(i_{n'}).$$

If  $x$  contains a middle palindrome of length exactly  $2k$ , then  $w(x)$  contains a middle palindrome of length  $2(1+2\varepsilon)^{k+1} \cdot n^{0.99}$ . Also, based on the properties of  $\nu$ , any non-middle centered palindrome in  $w(x)$  has length at most  $O(\sqrt{n})$ , which is less than  $n^{0.99}$  for  $n$  large enough. Since  $\lceil 2(1+2\varepsilon)^k \cdot n^{0.99} \rceil \cdot (1+\varepsilon) < (2(1+2\varepsilon)^k \cdot n^{0.99} + 1) \cdot (1+\varepsilon) < 2(1+2\varepsilon)^{k+1} \cdot n^{0.99}$ , value of  $k$  can be extracted from the answer of  $\mathcal{A}$ . Thus, if  $\mathcal{A}$  approximates the middle palindrome in  $w(x)$  with multiplicative error  $(1+\varepsilon)$  with probability  $1 - \frac{1}{n}$  using  $o(\frac{\log n}{\log(1+\varepsilon)} \log \sigma)$  bits of memory, we can construct a new algorithm  $\mathcal{A}'$  solving  $\text{midLPS}_\Sigma[2n']$  exactly with probability  $1 - \frac{1}{n} > 1 - \frac{1}{2n'}$  using

$$o(\frac{\log n}{\log(1+\varepsilon)} \log \sigma) + \log n \quad (7)$$

bits of memory. By Lemma 2.4 we get a lower bound

$$\begin{aligned} \gamma \cdot 2n' \log \min\{|\Sigma|, 2n'\} &= \frac{\gamma}{50} \cdot \frac{\log n}{\log(1+2\varepsilon)} \log \sigma - 2\gamma \log \sigma \\ &\geq \frac{\gamma}{100} \cdot \frac{\log n}{\log(1+2\varepsilon)} \log \sigma + \log n - 2\gamma \log \sigma \end{aligned} \quad (8)$$

(where the last inequality holds because of (6)). On the other hand, for large  $n$

$$\begin{aligned} \frac{\gamma}{100} \cdot \frac{\log n}{\log(1+2\varepsilon)} \log \sigma - 2\gamma \log \sigma + \log n &= \left( \frac{1}{100} \frac{\log n}{\log(1+2\varepsilon)} - 2 \right) \gamma \log \sigma + \log n \\ &= \Theta\left(\frac{\log n}{\log(1+\varepsilon)} \log \sigma\right) + \log n \end{aligned}$$

so (8) exceeds (7), a contradiction.  $\square$

### 3 Real-Time Algorithms

In this section we design real-time Monte Carlo algorithms within the space bounds matching the lower bounds from Sect. 2 up to a factor bounded by  $\log n$ . The algorithms make use of the hash function known as the *Karp-Rabin fingerprint* [11]. Let  $p$  be a fixed prime from the range  $[n^{3+\alpha}, n^{4+\alpha}]$  for some  $\alpha > 0$ , and  $r$  be a fixed integer randomly chosen from  $\{1, \dots, p-1\}$ . For a string  $S$ , its forward hash and reversed hash are defined, respectively, as

$$\phi^F(S) = \left( \sum_{i=1}^n S[i] \cdot r^i \right) \bmod p \quad \text{and} \quad \phi^R(S) = \left( \sum_{i=1}^n S[i] \cdot r^{n-i+1} \right) \bmod p.$$

Clearly, the forward hash of a string coincides with the reversed hash of its reversal. Thus, if  $u$  is a palindrome, then  $\phi^F(u) = \phi^R(u)$ . The converse is also true modulo the (improbable) collisions of hashes, because for two strings  $u \neq v$  of length  $m$ , the probability that  $\phi^F(u) = \phi^F(v)$  is at most  $m/p$ . This property allows one to detect palindromes with high probability by comparing hashes. (This approach is somewhat simpler than the one of [2]; in particular, we do not need “fingerprint pairs” used there.) In particular, a real-time algorithm makes  $\mathcal{O}(n)$  comparisons and thus faces a collision with probability  $\mathcal{O}(n^{-1-\alpha})$  by the choice of  $p$ . All further considerations assume that no collisions happen. For an input stream  $S$ , we denote  $F^F(i, j) = \phi^F(S[i..j])$  and  $F^R(i, j) = \phi^R(S[i..j])$ . The next observation is quite important.

**Proposition 3.1** ([3]). *The following equalities hold:*

$$\begin{aligned} F^F(i, j) &= r^{-(i-1)} (F^F(1, j) - F^F(1, i-1)) \bmod p, \\ F^R(i, j) &= F^R(1, j) - r^{j-i+1} F^R(1, i-1) \bmod p. \end{aligned}$$

Let  $I(i)$  denote the tuple  $(i, F^F(1, i-1), F^R(1, i-1), r^{-(i-1)} \bmod p, r^i \bmod p)$ . The proposition below is immediate from definitions and Proposition 3.1.

**Proposition 3.2.** 1) *Given  $I(i)$  and  $S[i]$ , the tuple  $I(i+1)$  can be computed in  $\mathcal{O}(1)$  time.*  
 2) *Given  $I(i)$  and  $I(j+1)$ , the string  $S[i..j]$  can be checked for being a palindrome in  $\mathcal{O}(1)$  time.*

### 3.1 Additive Error

**Theorem 3.3.** *There is a real-time Monte Carlo algorithm solving the problem  $\text{LPS}(S)$  with the additive error  $E = E(n)$  using  $\mathcal{O}(n/E)$  space, where  $n = |S|$ .*

First we present a simple (and slow) algorithm which solves the posed problem, i.e., finds in  $S$  a palindrome of length  $\ell(S) \geq L(S) - E$ , where  $L(S)$  is the length of the longest palindrome in  $S$ . Later this algorithm will be converted into a real-time one. We store the sets  $I(j)$  for some values of  $j$  in a doubly-linked list  $SP$  in the decreasing order of  $j$ 's. The longest palindrome currently found is stored as a pair  $answer = (pos, len)$ , where  $pos$  is its initial position and  $len$  is its length. Let  $t_E = \lfloor \frac{E}{2} \rfloor$ .

In Algorithm ABasic we add  $I(j)$  to the list  $SP$  for each  $j$  divisible by  $t_E$ . This allows us to check for palindromicity, at  $i$ th iteration, all factors of the form  $S[kt_E..i]$ . We assume throughout the section that at the beginning of  $i$ th iteration the value  $I(i)$  is stored in a variable  $I$ .

---

**Algorithm 1** : Algorithm ABasic,  $i$ th iteration

---

- 1: **if**  $i \bmod t_E = 0$  **then**
  - 2:     add  $I$  to the beginning of  $SP$
  - 3: read  $S[i]$ ; compute  $I(i+1)$  from  $I$ ;  $I \leftarrow I(i+1)$
  - 4: **for** all elements  $v$  of  $SP$  **do**
  - 5:     **if**  $S[v.i..i]$  is a palindrome and  $answer.len < i-v.i+1$  **then**
  - 6:          $answer \leftarrow (v.i, i-v.i+1)$
- 

**Proposition 3.4.** *Algorithm ABasic finds in  $S$  a palindrome of length  $\ell(S) \geq L(S) - E$  using  $\mathcal{O}(n/E)$  time per iteration and  $\mathcal{O}(n/E)$  space.*

*Proof.* Both the time and space bounds arise from the size of the list  $SP$ , which is bounded by  $n/t_E = \mathcal{O}(n/E)$ ; the number of operations per iteration is proportional to this size due to Proposition 3.2. Now let  $S[i..j]$  be a longest palindrome in  $S$ . Let  $k = \lceil \frac{i}{t_E} \rceil t_E$ . Then  $i \leq k < i + t_E$ . At the  $k$ th iteration,  $I(k)$  was added to  $SP$ ; then the palindrome  $S[k..j-(k-i)]$  was found at the iteration  $j - (k - i)$ . Its length is

$$j - (k - i) - k + 1 = j - i - 2(k - i) + 1 > (j - i + 1) - 2t_E = L(S) - 2 \left\lfloor \frac{E}{2} \right\rfloor \geq L(S) - E,$$

as required. □



The resource to speed up Algorithm ABasic stems from the following

**Lemma 3.5.** *During one iteration, the length  $answer.len$  increases by at most  $2 \cdot t_E$ .*

*Proof.* Let  $S[j..i]$  be the longest palindrome found at the  $i$ th iteration. If  $i - j + 1 \leq 2t_E$  then the statement is obviously true. Otherwise the palindrome  $S[j+t_E..i-t_E]$  of length  $i - j + 1 - 2t_E$  was found before (at the  $(i-t_E)$ th iteration), and the statement holds again.  $\square$

Lemma 3.5 implies that at each iteration  $SP$  contains only two elements that can increase  $answer.len$ . Hence we get the following Algorithm A.

---

**Algorithm 2 :** Algorithm A,  $i$ th iteration

---

```

1: if  $i \bmod t_E = 0$  then
2:   add  $I$  to the beginning of  $SP$ 
3:   if  $i = t_E$  then
4:      $sp \leftarrow first(SP)$ 
5: read  $S[i]$ ; compute  $I(i+1)$  from  $I$ ;  $I \leftarrow I(i+1)$ 
6:  $sp \leftarrow previous(sp)$  ▷ if exists
7: while  $i - sp.i + 1 \leq answer.len$  and  $(sp \neq last(SP))$  do
8:    $sp \leftarrow next(sp)$ 
9: for all existing  $v$  in  $\{sp, next(sp)\}$  do
10:  if  $S[v.i..i]$  is a palindrome and  $answer.len < i - v.i + 1$  then
11:     $answer \leftarrow (v.i, i - v.i + 1)$ 

```

---

Due to Lemma 3.5, the cycle at lines 9–11 of Algorithm A computes the same sequence of values of  $answer$  as the cycle at lines 4–6 of Algorithm ABasic. Hence it finds a palindrome of required length by Proposition 3.4. Clearly, the space used by the two algorithms differs by a constant. To prove that an iteration of Algorithm A takes  $\mathcal{O}(1)$  time, it suffices to note that the cycle in lines 7–8 performs at most two iterations. Theorem 3.3 is proved.

## 3.2 Multiplicative Error for $\varepsilon \leq 1$

**Theorem 3.6.** *There is a real-time Monte Carlo algorithm solving the problem  $LPS(S)$  with multiplicative error  $\varepsilon = \varepsilon(n) \in (0, 1]$  using  $\mathcal{O}\left(\frac{\log(n\varepsilon)}{\varepsilon}\right)$  space, where  $n = |S|$ .*

As in the previous section, we first present a simpler algorithm MBasic with non-linear working time and then upgrade it to a real-time algorithm. The algorithm must find a palindrome of length  $\ell(S) \geq \frac{L(S)}{1+\varepsilon}$ . The next lemma is straightforward.

**Lemma 3.7.** *If  $\varepsilon \in (0, 1]$ , the condition  $\ell(S) \geq L(S)(1 - \varepsilon/2)$  implies  $\ell(S) \geq \frac{L(S)}{1+\varepsilon}$ .*

We set  $q_\varepsilon = \lceil \log \frac{2}{\varepsilon} \rceil$ . The main difference in the construction of algorithms with the multiplicative and additive error is that here all sets  $I(i)$  are added to the list  $SP$ , but then, after a certain number of steps, are deleted from it. The number of iterations the set  $I(i)$  is stored in  $SP$  is determined by the time-to-live function  $tll(i)$  defined below. This function is responsible for both the correctness of the algorithm and the space bound.

Let  $\beta(i)$  be the position of the rightmost 1 in the binary representation of  $i$  (the position 0 corresponds to the least significant bit). We define

$$tll(i) = 2^{q_\varepsilon + 2 + \beta(i)}. \quad (9)$$

The definition is illustrated by Fig. 1. Next we state a few properties of the list  $SP$ .

**Lemma 3.8.** *For any integers  $a \geq 1$  and  $b \geq 0$ , there exists a unique integer  $j \in [a, a + 2^b)$  such that  $tll(j) \geq 2^{q_\varepsilon + 2 + b}$ .*

---

**Algorithm 3** : Algorithm MBasic,  $i$ th iteration

---

```

1: add  $I$  to the beginning of  $SP$ 
2: for all  $v$  in  $SP$  do
3:   if  $v.i + ttl(v.i) = i$  then
4:     delete  $v$  from  $SP$ 
5: read  $S[i]$ ; compute  $I(i+1)$  from  $I$ ;  $I \leftarrow I(i+1)$ 
6: for all  $v$  in  $SP$  do
7:   if  $S[v.i..i]$  is a palindrome and  $answer.len < i-v.i+1$  then
8:      $answer \leftarrow (v.i, i-v.i+1)$ 

```

---

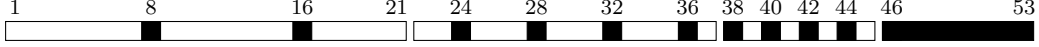


Figure 1: The state of the list  $SP$  after the iteration  $i = 53$  ( $q_\varepsilon = 1$  is assumed). Black squares indicate the numbers  $j$  for which  $I(j)$  is currently stored. For example, (9) implies  $ttl(28) = 2^{1+2+2} = 32$ , so  $I(28)$  will stay in  $SP$  until the iteration  $28 + 32 = 60$ .

*Proof.* By (9),  $ttl(j) \geq 2^{q_\varepsilon+2+b}$  if and only if  $\beta(j) \geq b$ , i.e.,  $j$  is divisible by  $2^b$  by the definition of  $\beta$ . Among any  $2^b$  consecutive integers, exactly one has this property.  $\square$

Figure 1 shows the partition of the range  $(0, i]$  into intervals having lengths that are powers of 2 (except for the leftmost interval). In general, this partition consists of the following intervals, right to left:

$$(i - 2^{q_\varepsilon+2}, i], (i - 2^{q_\varepsilon+3}, i - 2^{q_\varepsilon+2}], \dots, (i - 2^m, i - 2^{m-1}], (0, i - 2^m], \text{ where } m = \left\lceil \log \frac{n}{2^{q_\varepsilon+2}} \right\rceil - 1. \quad (10)$$

Lemma 3.8 and (9) imply the following lemma on the distribution of the elements of  $SP$ .

**Lemma 3.9.** *After each iteration, the first interval (resp., the last interval; each of the remaining intervals) in (10) contains  $2^{q_\varepsilon+2}$  (resp., at most  $2^{q_\varepsilon+1}$ ; exactly  $2^{q_\varepsilon+1}$ ) elements of the list  $SP$ .*

The number of the intervals in (10) is  $\mathcal{O}(\log(n\varepsilon))$ , so from Lemma 3.9 and the definition of  $q_\varepsilon$  we have the following.

**Lemma 3.10.** *After each iteration, the size of the list  $SP$  is  $\mathcal{O}\left(\frac{\log(n\varepsilon)}{\varepsilon}\right)$ .*

**Proposition 3.11.** *Algorithm MBasic finds a palindrome of length  $\ell(S) \geq \frac{L(S)}{1+\varepsilon}$  using  $\mathcal{O}\left(\frac{\log(n\varepsilon)}{\varepsilon}\right)$  time per iteration and  $\mathcal{O}\left(\frac{\log(n\varepsilon)}{\varepsilon}\right)$  space.*

*Proof.* Both the time per iteration and the space are dominated by the size of the list  $SP$ . Hence the required complexity bounds follow from Lemma 3.10. For the proof of correctness, let  $S[i..j]$  be a palindrome of length  $L(S)$ . Further, let  $d = \lfloor \log L(S) \rfloor$ .

If  $d < q_\varepsilon + 2$ , the palindrome  $S[i..j]$  will be found exactly, because  $I(i)$  is in  $SP$  at the  $j$ th iteration:

$$i + ttl(i) \geq i + 2^{q_\varepsilon+2} \geq i + 2^{d+1} > i + L(S) > j.$$

Otherwise, by Lemma 3.8 there exists a unique  $k \in [i, i + 2^{d-q_\varepsilon-1})$  such that  $ttl(k) \geq 2^{d+1}$ . Hence at the iteration  $j - (k - i)$  the palindrome  $S[i + (k - i) .. j - (k - i)]$  will be found, because  $I(k)$  is in  $SP$  at this iteration:

$$k + ttl(k) \geq i + ttl(k) \geq i + 2^{d+1} > j \geq j - (k - i).$$

The length of this palindrome satisfies the requirement of the proposition:

$$j - (k - i) - (i + (k - i)) + 1 = L(S) - 2(k - i) \geq L(S) - 2^{d-q_\varepsilon} \geq L(S) - \frac{L(S)}{2^{q_\varepsilon}} \geq L(S) \left(1 - \frac{\varepsilon}{2}\right).$$

The reference to Lemma 3.7 finishes the proof.  $\square$

Now we speed up Algorithm MBasic. It has two slow parts: deletions from the list  $SP$  and checks for palindromes. Lemmas 3.12 and 3.13 show that, similar to Sect. 3.1,  $\mathcal{O}(1)$  checks are enough at each iteration.

**Lemma 3.12.** *Suppose that at some iteration the list  $SP$  contains consecutive elements  $I(d), I(c), I(b), I(a)$ . Then  $b - a \leq d - b$ .*

*Proof.* Let  $j$  be the number of the considered iteration. Note that  $a < b < c < d$ . Consider the interval in (10) containing  $a$ . If  $a \in (j - 2^{q_\varepsilon+2}, j]$ , then  $b - a = 1$  and  $d - b = 2$ , so the required inequality holds. Otherwise, let  $a \in (j - 2^{q_\varepsilon+2+x}, j - 2^{q_\varepsilon+2+x-1}]$ . Then by (9)  $\beta(a) \geq x$ ; moreover, any  $I(k)$  such that  $a < k \leq j$  and  $\beta(k) \geq x$  is in  $SP$ . Hence,  $b - a \leq 2^x$ . By Lemma 3.9 each interval, except for the leftmost one, contains at least  $2^{q_\varepsilon+1} \geq 4$  elements. Thus each of the numbers  $b, c, d$  belongs either to the same interval as  $a$  or to the previous interval  $(j - 2^{q_\varepsilon+2+x-1}, j - 2^{q_\varepsilon+2+x-2}]$ . Again by (9) we have  $\beta(b), \beta(c), \beta(d) \geq x - 1$ . So  $c - b, d - c \geq 2^{x-1}$ , whence the result.  $\square$

We call an element  $I(a)$  of  $SP$  *valuable at  $i$ th iteration* if  $i - a + 1 > \text{answer.len}$  and  $S[a..i]$  can be a palindrome. (That is, Algorithm MBasic does not store enough information to predict that the condition in its line 7 is false for  $v = I(a)$ .)

**Lemma 3.13.** *At each iteration,  $SP$  contains at most three valuable elements. Moreover, if  $I(d'), I(d)$  are consecutive elements of  $SP$  and  $i - d' < \text{answer.len} \leq i - d$ , where  $i$  is the number of the current iteration, then the valuable elements are consecutive in  $SP$ , starting with  $I(d)$ .*

*Proof.* Let  $d$  be as in the condition of the lemma. If  $I(d)$  is followed in  $SP$  by at most two elements, we are done. If it is not the case, let the next three elements be  $I(c), I(b)$ , and  $I(a)$ , respectively. If  $S[a..i]$  is a palindrome then  $S[a+(b-a)..i-(b-a)]$  is also a palindrome. At the iteration  $i-(b-a)$  the tuple  $I(b)$  was in  $SP$ , so this palindrome was found. Hence, at the  $i$ th iteration the value  $\text{answer.len}$  is at least the length of this palindrome, which is  $i - a + 1 - 2(b - a)$ . By Lemma 3.12,  $b - a \leq d - b$ , implying  $\text{answer.len} \geq i - a + 1 - (b - a) - (d - b) = i - d + 1$ . This inequality contradicts the definition of  $d$ ; hence,  $S[a..i]$  is not a palindrome. By the same argument, the elements following  $I(a)$  in  $SP$  do not produce palindromes as well. Thus, only the elements  $I(d), I(c), I(b)$  are valuable.  $\square$

Now we turn to deletions. The function  $\text{ttl}(x)$  has the following nice property.

**Lemma 3.14.** *The function  $x \rightarrow x + \text{ttl}(x)$  is injective.*

*Proof.* Note that  $\beta(x + \text{ttl}(x)) = \beta(x)$  from the definition of  $\text{ttl}$ . Hence the equality  $x + \text{ttl}(x) = y + \text{ttl}(y)$  implies  $\beta(x) = \beta(y)$ , then  $\text{ttl}(x) = \text{ttl}(y)$  by (9), and finally  $x = y$ .  $\square$

Lemma 3.14 implies that at most one element is deleted from  $SP$  at each iteration. To perform this deletion in  $\mathcal{O}(1)$  time, we need an additional data structure. By  $BS(x)$  we denote a linked list of maximal segments of 1's in the binary representation of  $x$ . For example, the binary representation of  $x = 12345$  and  $BS(x)$  are as follows:

13	12	11	10	9	8	7	6	5	4	3	2	1	0
1	1	0	1	0	0	0	0	1	1	1	0	0	1

$BS(12345) = \{[0, 0], [3, 5], [10, 10], [12, 13]\}$

Clearly,  $BS(x)$  uses  $\mathcal{O}(\log x)$  space.

**Lemma 3.15.** *Both  $\beta(x)$  and  $BS(x + 1)$  can be obtained from  $BS(x)$  in  $\mathcal{O}(1)$  time.*

*Proof.* The first number in  $BS(x)$  is  $\beta(x)$ . Let us construct  $BS(x + 1)$ . Let  $[a, b]$  be the first segment in  $BS(x)$ . If  $a > 2$ , then  $BS(x + 1) = [0, 0] \cup BS(x)$ . If  $a = 1$ , then  $BS(x + 1) = [0, b] \cup (BS(x) \setminus [1, b])$ . Now let  $a = 0$ . If  $BS(x) = \{[0, b]\}$  then  $BS(x + 1) = \{[b+1, b+1]\}$ . Otherwise let the second segment in  $BS(x)$  be  $[c, d]$ . If  $c > b + 2$ , then  $BS(x + 1) = [b+1, b+1] \cup (BS(x) \setminus [0, b])$ . Finally, if  $c = b + 2$ , then  $BS(x + 1) = [b+1, d] \cup (BS(x) \setminus \{[0, b], [c, d]\})$ .  $\square$

Thus, if we support one list  $BS$  which is equal to  $BS(i)$  at the end of the  $i$ th iteration, we have  $\beta(i)$ . If  $I(a)$  should be deleted from  $SP$  at this iteration, then  $\beta(a) = \beta(i)$  (see Lemma 3.14). The following lemma is trivial.

**Lemma 3.16.** *If  $a < b$  and  $tll(a) = tll(b)$ , then  $I(a)$  is deleted from  $SP$  before  $I(b)$ .*

By Lemma 3.16, the information about the positions with the same  $tll$  (in other words, with the same  $\beta$ ) are added to and deleted from  $SP$  in the same order. Hence it is possible to keep a queue  $QU(x)$  of the pointers to all elements of  $SP$  corresponding to the positions  $j$  with  $\beta(j) = x$ . These queues constitute the last ingredient of our real-time Algorithm M.

---

**Algorithm 4 :** Algorithm M,  $i$ th iteration

---

```

1: add  $I$  to the beginning of  $SP$ 
2: if  $i = 1$  then
3:    $sp \leftarrow first(SP)$ 
4: compute  $BS[i]$  from  $BS$ ;  $BS \leftarrow BS[i]$ ; compute  $\beta(i)$  from  $BS$ 
5: if  $QU(\beta(i))$  is not empty then
6:    $v \leftarrow$  element of  $SP$  pointed by  $first(QU(\beta(i)))$ 
7:   if  $v = sp$  then
8:      $sp \leftarrow next(sp)$ 
9:   delete  $v$ ; delete  $first(QU(\beta(i)))$ 
10: add pointer to  $first(SP)$  to  $QU(\beta(i))$ 
11: read  $S[i]$ ; compute  $I(i+1)$  from  $I$ ;  $I \leftarrow I(i+1)$ 
12:  $sp \leftarrow previous(sp)$  ▷ if exists
13: while  $i - sp.i + 1 \leq answer.len$  and  $sp \neq last(SP)$  do
14:    $sp \leftarrow next(sp)$ 
15: for all existing  $v$  in  $\{sp, next(sp), next(next(sp))\}$  do
16:   if  $S[v.i..i]$  is a palindrome and  $answer.len < i - v.i + 1$  then
17:      $answer \leftarrow (v.i, i - v.i + 1)$ 

```

---

*Proof of Theorem 3.6.* After every iteration, Algorithm M has the same list  $SP$  (see Fig. 1) as Algorithm MBasic, because these algorithms add and delete the same elements. Due to Lemma 3.13, Algorithm M returns the same answer as Algorithm MBasic. Hence by Proposition 3.11 Algorithm M finds a palindrome of required length. Further, Algorithm M supports the list  $BS$  of size  $\mathcal{O}(\log n)$  and the array  $QU$  containing  $\mathcal{O}(\log n)$  queues of total size equal to the size of  $SP$ . Hence, it uses  $\mathcal{O}(\frac{\log(n\varepsilon)}{\varepsilon})$  space in total by Lemma 3.10. The cycle in lines 13–14 performs at most three iterations. Indeed, let  $z$  be the value of  $sp$  after the previous iteration. Then this cycle starts with  $sp = previous(z)$  (or with  $sp = z$  if  $z$  is the first element of  $SP$ ) and ends with  $sp = next(next(z))$  at the latest. By Lemma 3.15, both  $BS(i)$  and  $\beta(i)$  can be computed in  $\mathcal{O}(1)$  time. Therefore, each iteration takes  $\mathcal{O}(1)$  time.  $\square$

**Remark** Since for  $n^{-0.99} \leq \varepsilon \leq 1$  the classes  $\mathcal{O}(\frac{\log n}{\log(1+\varepsilon)})$  and  $\mathcal{O}(\frac{\log(n\varepsilon)}{\varepsilon})$  coincide, Algorithm M uses space within a  $\log n$  factor from the lower bound of Theorem 2.7. Furthermore, for an arbitrarily slowly growing function  $\varphi$  Algorithm M uses  $o(n)$  space whenever  $\varepsilon = \frac{\varphi(n)}{n}$ .

### 3.3 Multiplicative Error for $\varepsilon > 1$

**Theorem 3.17.** *There is a real-time Monte Carlo algorithm solving the problem  $LPS(S)$  with multiplicative error  $\varepsilon = \varepsilon(n) \in (1, n]$  using  $\mathcal{O}(\frac{\log(n)}{\log(1+\varepsilon)})$  space, where  $n = |S|$ .*

We transform Algorithm M into real-time Algorithm M' which solves  $\text{LPS}(S)$  with the multiplicative error  $\varepsilon > 1$  using  $\mathcal{O}\left(\frac{\log n}{\log(1+\varepsilon)}\right)$  space. The basic idea of transformation is to replace all binary representations with those in base proportional to  $1 + \varepsilon$ , and thus shrink the size of the lists  $SP$  and  $BS$ .

First, we assume without loss of generality that  $\varepsilon \geq 7$ , as otherwise we can fix  $\varepsilon = 1$  and apply Algorithm M. Fix  $k \leq \frac{1}{2}(1 + \varepsilon)$  as the largest such even integer (in particular,  $k \geq 4$ ). Let  $\beta'(i)$  be the position of the rightmost non-zero digit in the  $k$ -ary representation of  $i$ . We define

$$ttl'(i) = \begin{cases} \frac{9}{2} \cdot k^{\beta'(i)} & \text{if } \beta'(i) > 0, \\ 4 & \text{otherwise.} \end{cases} \quad (11)$$

The list  $SP'$  is the analog of the list  $SP$  from Sect. 3.2. It contains, after  $i$ th iteration, the tuples  $I(j)$  for all positions  $j \leq i$  such that  $j + ttl'(j) > i$ . Similar to (10), we partition the range  $(0; i]$  into intervals and then count the elements of  $SP'$  in these intervals. The intervals are, right to left,

$$(i - 4, i], (i - \frac{9}{2}k, i - 4], (i - \frac{9}{2}k^2, i - \frac{9}{2}k], \dots, (i - \frac{9}{2}k^m, i - \frac{9}{2}k^{m-1}], (0, i - \frac{9}{2}k^m]. \quad (12)$$

For convenience, we enumerate the intervals starting with 0.

**Lemma 3.18.** *Each interval in (12) contains at most 5 elements of  $SP'$ . Each of the intervals  $0, \dots, m$  contains at least 3 elements of  $SP'$ .*

*Proof.* The 0th interval contains exactly 4 elements. For any  $j = 1, \dots, m+1$ , an element  $x$  of the  $j$ th interval is in  $SP'$  if and only if its position is divisible by  $k^j$ ; see (11). The length of this interval is less than  $\frac{9}{2}k^j$ , giving us upper bound of  $\lceil \frac{9}{2} \rceil = 5$  elements. Similarly, if  $j \neq m+1$ , the  $j$ th interval has the length  $\frac{9}{2}k^j - \frac{9}{2}k^{j-1}$  and thus contains at least  $\lfloor \frac{9}{2} \frac{k-1}{k} \rfloor$  elements of  $SP'$ . Since  $k \geq 4$ , the claim follows.  $\square$

Next we modify Algorithm MBasic, replacing  $ttl$  by  $ttl'$  and  $SP$  by  $SP'$ .

**Proposition 3.19.** *Modified Algorithm MBasic finds a palindrome of length  $\ell(S) \geq \frac{L(S)}{1+\varepsilon}$  using  $\mathcal{O}\left(\frac{\log n}{\log(1+\varepsilon)}\right)$  space.*

*Proof.* Let  $S[i..j]$  be a palindrome of length  $L(S)$ . Let  $d = \lfloor \log_k \frac{L(S)}{4} \rfloor$ . Without loss of generality we assume  $d \geq 0$ , as otherwise  $L(S) < 4 \leq ttl'(i)$  and the palindrome  $S[i..j]$  will be detected exactly. Since  $L(S) \geq 4k^d$ , let  $a_1 < a_2 < a_3 < a_4 < a_5$  be consecutive positions which are multiples of  $k^d$  (i.e.,  $\beta'(a_1), \dots, \beta'(a_5) \geq d$ ) such that  $a_2 \leq \frac{i+j}{2} < a_3$ . Then in particular  $i < a_1$ , and there is a palindrome  $S[a_1..(i+j-a_1)]$  such that  $a_3 \leq (i+j-a_1) < a_5$ . Since  $a_1 + ttl'(a_1) \geq a_5$ , this particular palindrome will be detected by the modified Algorithm MBasic; thus  $\ell(S) \geq a_3 - a_1 = 2k^d$ . However, we have  $L(S) < 4k^{d+1}$ , hence  $\frac{L(S)}{\ell(S)} < 2k \leq (1 + \varepsilon)$ .

Space complexity follows from bound on size of the list  $SP'$ , which is at most  $5 \lceil \frac{\log n}{\log k} \rceil = \mathcal{O}\left(\frac{\log n}{\log(1+\varepsilon)}\right)$ .  $\square$

To follow the framework from the case of  $\varepsilon \leq 1$ , we provide analogous speedup to the checks for palindromes. We adopt the same notion of an element valuable at  $i$ th iteration as in Sect. 3.2. First we need the following property, which is a more general analog of Lemma 3.12; an analog of Lemma 3.13 is then proved with its help.

**Lemma 3.20.** *Suppose that at some iteration the list  $SP'$  contains consecutive elements  $I(d), I(c)$ , and  $d \leq i - \text{answer.len}$ , where  $i$  is the number of the current iteration. Further, let  $I(a)$  be another element of  $SP'$  at this iteration and  $a < c$ . If  $c, d$  belong to the same interval of (12), then  $I(a)$  is not valuable.*

*Proof.* Let  $c, d$  belong to the  $j$ th interval. Thus they are divisible by  $k^j$  and  $d - c = k^j$ . Since  $a < c$ ,  $a$  is divisible by  $k^j$  as well. One of the numbers  $\frac{d+a}{2}, \frac{c+a}{2}$  is divisible by  $k^j$ ; take it as  $b$ . Let  $\delta = b - a$ . If  $S[a..i]$  is a palindrome, then  $S[b..i - \delta]$  is also a palindrome. Since at the  $i$ th iteration the left border of the  $j$ th interval was smaller than  $c$ , then at the  $(i - \delta)$ th iteration this border was smaller than  $b$ ; hence,  $I(b)$  was in  $SP'$  at that iteration, and the palindrome  $S[b..i - \delta]$  was found. Its length is

$$i - \delta - b + 1 = i + 1 + a - 2b \geq i + 1 + a - (d + a) \geq i - d + 1 > \text{answer.len},$$

which is impossible by the definition of  $\text{answer.len}$ . So  $S[a..i]$  is not a palindrome, and the claim follows.  $\square$

**Lemma 3.21.** *At each iteration,  $SP'$  contains at most three valuable elements. Moreover, if  $I(d'), I(d)$  are consecutive elements of  $SP'$  and  $i - d' < \text{answer.len} \leq i - d$ , where  $i$  is the number of the current iteration, then the valuable elements are consecutive in  $SP'$ , starting with  $I(d)$ .*

*Proof.* Let  $a < b < c < d$  be such that the elements  $I(d), I(c), I(b)$  are consecutive in  $SP'$  and  $I(a)$  belongs to  $SP'$ . Then either  $b, c$  or  $c, d$  are in the same interval of (12), and thus  $a$  is not valuable by Lemma 3.20.  $\square$

To complete the proof, we now turn to deletions, proving the following analog of Lemma 3.14.

**Lemma 3.22.** *The function  $h(x) = x + \text{ttl}'(x)$  maps at most two different values of  $x$  to the same value. Moreover, if  $h(x) = h(y)$  and  $\beta'(x) \geq \beta'(y)$ , then  $\beta'(x) = \beta'(h(x)) + 1$  and  $\beta'(y) = 0$ .*

*Proof.* Let  $h(x) = h(y)$ . If  $\beta'(x) = \beta'(y)$  then  $\text{ttl}'(x) = \text{ttl}'(y)$  by (11), implying  $x = y$ . Hence all preimages of  $h(x)$  have different values of  $\beta'$ . Assume  $\beta'(x) > \beta'(y)$ . Then we have, for some integer  $j$ ,  $x = j \cdot k^{\beta'(x)}$  and  $h(x) = (j + 4)k^{\beta'(x)} + \frac{k}{2} \cdot k^{\beta'(x)-1}$  by (11). Since  $k$  is even, we get  $\beta'(h(x)) = \beta'(x) - 1$ . If  $\beta'(y) > 0$ , we repeat the same argument and obtain  $\beta'(x) = \beta'(y)$ , contradicting our assumption. Thus  $\beta'(y) = 0$ . The claim now follows.  $\square$

We also define a list  $BS'(x)$ , which maintains an RLE encoding of the  $k$ -ary representation of  $x$ . The list  $BS'(x)$  has length  $\mathcal{O}(\frac{\log n}{\log k})$ , can be updated to  $BS'(x+1)$  in  $\mathcal{O}(1)$  time, and provides the value  $\beta'(x)$  in  $\mathcal{O}(1)$  time also (cf. Lemma 3.15). Further, Lemma 3.16 holds for the function  $\text{ttl}'$ , so we introduce the queues  $QU'(x)$  in the same way as the queues  $QU(x)$  in Sect. 3.2. Having all the ingredients, we present Algorithm M' which speeds up the modified Algorithm MBasic and thus proves Theorem 3.17. The only significant difference between Algorithm M and Algorithm M' is in the deletion of tuples from the list (compare lines 5–9 of Algorithm M against lines 5–15 of Algorithm M').

---

**Algorithm 5 :** Algorithm M',  $i$ th iteration

---

```

1: add  $I$  to the beginning of  $SP'$ 
2: if  $i = 1$  then
3:    $sp \leftarrow \text{first}(SP')$ 
4: compute  $BS'[i]$  from  $BS'$ ;  $BS' \leftarrow BS'[i]$ ; compute  $\beta(i)$  from  $BS'$ 
5: if  $QU'(\beta(i) + 1)$  is not empty then
6:    $v \leftarrow$  element of  $SP'$  pointed by  $\text{first}(QU'(\beta(i)))$ 
7:   if  $v.i + \text{ttl}'(v.i) = i$  then
8:     if  $v = sp$  then
9:        $sp \leftarrow \text{next}(sp)$ 
10:    delete  $v$ ; delete  $\text{first}(QU'(\beta(i) + 1))$ 
11:  $v \leftarrow$  element of  $SP'$  pointed by  $\text{first}(QU'(0))$ 
12: if  $v.i + \text{ttl}'(v.i) = i$  then
13:   if  $v = sp$  then
14:      $sp \leftarrow \text{next}(sp)$ 
15:   delete  $v$ ; delete  $\text{first}(QU'(0))$ 
16: add pointer to  $\text{first}(SP')$  to  $QU'(\beta(i))$ 
17: read  $S[i]$ ; compute  $I(i + 1)$  from  $I$ ;  $I \leftarrow I(i + 1)$ 
18:  $sp \leftarrow \text{previous}(sp)$  ▷ if exists
19: while  $i - sp.i + 1 \leq \text{answer.len}$  and  $sp \neq \text{last}(SP')$  do
20:    $sp \leftarrow \text{next}(sp)$ 
21: for all existing  $v$  in  $\{sp, \text{next}(sp), \text{next}(\text{next}(sp))\}$  do
22:   if  $S[v.i..i]$  is a palindrome and  $\text{answer.len} < i - v.i + 1$  then
23:      $\text{answer} \leftarrow (v.i, i - v.i + 1)$ 

```

---

### 3.4 The Case of Short Palindromes

A typical string contains only short palindromes, as Lemma 3.23 below shows (for more on palindromes in random strings, see [16]). Knowing this, it is quite useful to have a deterministic real-time algorithm which finds a longest palindrome exactly if it is “short”, otherwise reporting that it is “long”. The aim of this section is to describe such an algorithm (Theorem 3.24).

**Lemma 3.23.** *If an input stream  $S \in \Sigma^*$  is picked up uniformly at random among all strings of length  $n$ , where  $n \geq |\Sigma|$ , then for any positive constant  $c$  the probability that  $S$  contains a palindrome of length greater than  $\frac{2(c+1) \log n}{\log |\Sigma|}$  is  $\mathcal{O}(n^{-c})$ .*

*Proof.* A string  $S$  contains a palindrome of length greater than  $m$  if and only if  $S$  contains a palindrome of length  $m+1$  or  $m+2$ . The probability  $P$  of containing such a palindrome is less than the expected number  $M$  of palindromes of length  $m+1$  and  $m+2$  in  $S$ . A factor of  $S$  of length  $l$  is a palindrome with probability  $1/|\Sigma|^{\lfloor l/2 \rfloor}$ ; by linearity of expectation, we have

$$M = \frac{n - m}{|\Sigma|^{\lfloor (m+1)/2 \rfloor}} + \frac{n - m - 1}{|\Sigma|^{\lfloor (m+2)/2 \rfloor}}.$$

Substituting  $m = \frac{2(c+1) \log n}{\log |\Sigma|}$ , we get  $M = \mathcal{O}(n^{-c})$ , as required.  $\square$

**Theorem 3.24.** *Let  $m$  be a positive integer. There exists a deterministic real-time algorithm working in  $\mathcal{O}(m)$  space, which*

- solves  $\text{LPS}(S)$  exactly if  $L(S) < m$ ;
- finds a palindrome of length  $m$  or  $m+1$  as an approximated solution to  $\text{LPS}(S)$  if  $L(S) \geq m$ .

*Proof.* To prove Theorem 3.24, we present an algorithm based on the Manacher algorithm [14]. We add two features: work with a sliding window instead of the whole string to satisfy the space requirements and lazy computation to achieve real time. (The fact that the original Manacher algorithm admits a real-time version was shown by Galil [7]; we adjusted Galil’s approach to solve  $\text{LPS}$ .) The details follow.

We say that a palindromic factor  $S[i..j]$  has *center*  $\frac{i+j}{2}$  and *radius*  $\frac{j-i}{2}$ . Thus, odd-length (even-length) palindromes have integer (resp., half integer) centers and radiuses. This looks a bit weird, but allows one to avoid separate processing of these two types of palindromes. Manacher’s algorithm computes, in an online fashion, an array of maximal radiuses of palindromes centered at every position of the input string  $S$ . A variation, which outputs the length  $L$  of the longest palindrome in a string  $S$ , is presented as Algorithm EBasic below. This variation is similar to the one of [13]. Here,  $n$  stays for the length of the input processed so far,  $c$  is the center of the longest suffix-palindrome of the processed string. The array of radiuses  $Rad$  has length  $2n-1$  and its elements are indexed by all integers and half integers from the interval  $[1, n]$ . Initially,  $Rad$  is filled with zeroes. The left endmarker is added to the string for convenience. After each iteration, the following invariant holds: the element  $Rad[i]$  has got its true value if  $i < c$  and equals zero if  $i > c$ ; the value  $Rad[c] = n - c$  can increase at the next iteration. Note that the longest palindrome in  $S$  coincides with the longest suffix-palindrome of  $S[1..i]$  for some  $i$ . At the moment when the input stream ends, the algorithm has already found all such suffix-palindromes, so it can stop without filling the rest of the array  $Rad$ .

Note that  $n$  calls to `AddLetter` perform at most  $3n$  iterations of the cycle in lines 10–15 (each call performs the first iteration plus zero or more “additional” iterations; the value of  $c$  gets increased before each additional iteration and never decreases). So, Algorithm EBasic works in  $\mathcal{O}(n)$  time but not in real time; for example, reading the last letter of the string  $a^n b$  requires  $n$  iterations of the cycle.

By conditions of the theorem, we are not interested in palindromes of length  $> m+1$ . Thus, processing a suffix-palindrome of length  $m$  or  $m+1$  we assume that the symbol comparison in line 12 fails. So the procedure `AddLetter` needs no access to  $S[i]$  or  $Rad[i]$  whenever  $i < n - m$ . Hence we store only recent values of  $S$  and  $Rad$  and use circular arrays  $CS$  and  $CRad$  of size  $\mathcal{O}(m)$  for this purpose. For example, the symbol  $S[n-i]$  is stored in  $CS[(n-i) \bmod (m+1)]$  during  $m+1$  successful iterations of the outer cycle (lines 3–6), and then is replaced by  $S[n-i+m+1]$ ; the same scheme applies to the array  $Rad$ . In this way, all elements of  $S$  and  $Rad$ , needed by Algorithm EBasic, are accessible in constant time. Further, we define

---

**Algorithm 6** : Algorithm EBasic,  $i$ th iteration

---

```
1: procedure MANACHER
2:    $c \leftarrow 1$ ;  $L \leftarrow 1$ ;  $n \leftarrow 1$ ;  $S[0] \leftarrow \text{'\#}'$ 
3:   while not (end of input) do
4:     read( $S[n + 1]$ ); ADDLETTER( $S[n + 1]$ )
5:     if  $2 * \text{Rad}[c] + 1 > L$  then
6:        $L \leftarrow 2 * \text{Rad}[c] + 1$ 
7:   return  $L$ 
8: procedure ADDLETTER( $a$ )
9:    $s \leftarrow c$ 
10:  while  $c < n + 1$  do
11:     $\text{Rad}[c] \leftarrow \min(\text{Rad}[2 * s - c], n - c)$ 
12:    if  $c + \text{Rad}[c] = n$  and  $S[c - \text{Rad}[c] - 1] = a$  then
13:       $\text{Rad}[c] \leftarrow \text{Rad}[c] + 1$ 
14:      break ▷ longest suf-pal of  $S[1..n + 1]$  is found
15:       $c \leftarrow c + 0.5$  ▷ next candidate for the center
16:   $n \leftarrow n + 1$ 
```

---

a queue  $Q$  of size  $q$  for lazy computations; it contains symbols that are read from the input and await processing.

Now we describe real-time Algorithm E. It reads input symbols to  $Q$  and stops when the end of the input is reached. After reading a symbol, it runs procedure Manacher, requiring the latter to pause after three “inner iterations” (of the cycle in lines 10–15). The procedure reads symbols from  $Q$ ; if it tries to read from the empty queue, it pauses. When the procedure is called next time, it resumes from the moment it was stopped. Algorithm E returns the value of  $L$  after the last call to procedure Manacher.

To analyze Algorithm E, consider the value  $X = q + n - c$  after some iteration (clearly, this iteration has number  $q + n$ ) and look at the evolution of  $X$  over time. Let  $\Delta f$  denote the variation of the quantity  $f$  at one iteration. Note that  $\Delta(q + n) = 1$ . Let us describe  $\Delta X$ . First assume that the iteration contains three inner iterations. Then  $\Delta n = 0, 1, 2$  or  $3$  and, respectively,  $\Delta c = 1.5, 1, 0.5$  or  $0$ . Hence

$$\Delta X = 1 - \Delta c = 1 + (\Delta n - 3)/2 = 1 - (1 - \Delta q - 3)/2 = -(\Delta q)/2.$$

If the number of inner iterations is one or two, then  $q$  becomes zero (and was 0 or 1 before this iteration); hence  $\Delta n = 1 - \Delta q \geq 1$ . Then  $\Delta c \leq 0.5$  and finally  $\Delta X > 0$ . From these conditions on  $\Delta X$  it follows that

- (\*) if the current value of  $q$  is positive, then the current value of  $X$  is less than the value of  $X$  at the moment when  $q$  was zero for the last time.

Let  $X'$  be the previous value of  $X$  mentioned in (\*). Since the difference  $n - c$  does not exceed the radius of some palindrome,  $X' \leq m/2$ . Since  $q \leq X < X'$ , the queue  $Q$  uses  $\mathcal{O}(m)$  space. Therefore the same space bound applies to Algorithm E.

It remains to prove that Algorithm E returns the same number  $L$  as Algorithm EBasic with a sliding window, in spite of the fact that Algorithm E stops earlier (the content of  $Q$  remains unprocessed by procedure Manacher). Suppose that Algorithm E stops with  $q > 0$  after  $n$  iterations. Then the longest palindrome that could be found by processing the symbols in  $Q$  has the radius  $X = n + q - c$ . Now consider the iteration mentioned in (\*) and let  $n'$  and  $c'$  be the values of  $n$  and  $c$  after it; so  $X' = n' - c'$ . Since  $q$  was zero after that iteration, procedure Manacher read the symbol  $S[n']$  during it; hence, it tried to extend a suffix-palindrome of  $S[1..n' - 1]$  with the center  $c'' \leq c'$ . If this extension was successful, then a palindrome of radius at least  $X'$  was found. If it was unsuccessful, then  $c' \geq c'' + 1/2$  and hence  $S[1..n' - 1]$  has a suffix-palindrome of length at least  $X' - 1/2$ . Thus, a palindrome of length  $X \leq X' - 1/2$  is not longer than a longest palindrome seen before, and processing the queue cannot change the value of  $L$ . Thus, Algorithm E is correct.



The center and the radius of the longest palindrome in  $S$  can be updated each time the inequality in line 6 of procedure Manacher holds. Theorem 3.24 is proved.  $\square$

**Remark** Lemma 3.23 and Theorem 3.24 show a practical way to solve LPS. Algorithm E is fast and lightweight ( $2m$  machine words for the array *Rad*,  $m$  symbols in the sliding window and at most  $m$  symbols in the queue; compare to 17 machine words per one tuple  $I(i)$  in the Monte Carlo algorithms). So it makes direct sense to run Algorithm M and Algorithm E, both in  $\mathcal{O}(\log n)$  space, in parallel. Then either Algorithm E will give an exact answer (which happens with high probability if the input stream is a “typical” string) or both algorithms will produce approximations: one of fixed length and one with an approximation guarantee (modulo the hash collision).

## References

- [1] A. Apostolico, D. Breslauer, and Z. Galil. Parallel detection of all palindromes in a string. *Theoret. Comput. Sci.*, 141:163–173, 1995.
- [2] P. Berenbrink, F. Ergün, F. Mallmann-Trenn, and E. Sadeqi Azer. Palindrome recognition in the streaming model. In *STACS 2014*, volume 25 of *LIPIcs*, pages 149–161, Germany, 2014. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, Dagstuhl Publishing.
- [3] D. Breslauer and Z. Galil. Real-time streaming string-matching. In *Combinatorial Pattern Matching*, volume 6661 of *LNCS*, pages 162–172, Berlin, 2011. Springer.
- [4] R. Clifford, A. Fontaine, E. Porat, B. Sach, and T. A. Starikovskaya. Dictionary matching in a stream. In *ESA 2015*, volume 9294 of *LNCS*, pages 361–372. Springer, 2015.
- [5] R. Clifford, A. Fontaine, E. Porat, B. Sach, and T. A. Starikovskaya. The  $k$ -mismatch problem revisited. In *SODA 2016*, pages 2039–2052. SIAM, 2016.
- [6] F. Ergün, H. Jowhari, and M. Saglam. Periodicity in streams. In *RANDOM 2010*, volume 6302 of *LNCS*, pages 545–559. Springer, 2010.
- [7] Z. Galil. Real-time algorithms for string-matching and palindrome recognition. In *Proc. 8th annual ACM symposium on Theory of computing (STOC’76)*, pages 161–173, New York, USA, 1976. ACM.
- [8] Z. Galil and J. Seiferas. A linear-time on-line recognition algorithm for “Palstar”. *J. ACM*, 25:102–111, 1978.
- [9] P. Gawrychowski, F. Manea, and D. Nowotka. Testing generalised freeness of words. In *STACS 2014*, volume 25 of *LIPIcs*, pages 337–349. Dagstuhl Publishing, 2014.
- [10] M. Jalsenius, B. Porat, and B. Sach. Parameterized matching in the streaming model. In *STACS 2013*, volume 20 of *LIPIcs*, pages 400–411. Dagstuhl Publishing, 2013.
- [11] R. Karp and M. Rabin. Efficient randomized pattern matching algorithms. *IBM Journal of Research and Development*, 31:249–260, 1987.
- [12] D. E. Knuth, J. Morris, and V. Pratt. Fast pattern matching in strings. *SIAM J. Comput.*, 6:323–350, 1977.
- [13] D. Kosolobov, M. Rubinchik, and A. M. Shur. Finding distinct subpalindromes online. In *Proc. Prague Stringology Conference. PSC 2013*, pages 63–69. Czech Technical University in Prague, 2013.
- [14] G. Manacher. A new linear-time on-line algorithm finding the smallest initial palindrome of a string. *J. ACM*, 22(3):346–351, 1975.

- [15] B. Porat and E. Porat. Exact and approximate pattern matching in the streaming model. In *FOCS 2009*, pages 315–323. IEEE Computer Society, 2009.
- [16] M. Rubinchik and A. M. Shur. The number of distinct subpalindromes in random words. *Fundamenta Informaticae*, 145:371–384, 2016.
- [17] A. C.-C. Yao. Probabilistic computations: Toward a unified measure of complexity (extended abstract). In *FOCS 1977*, pages 222–227. IEEE Computer Society, 1977.